The Use of Ranked Set Sampling with Balanced Acceptance Sampling

Blair Robertson[1], Jennifer Brown[1], Trent McDonald[2]

[1] School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
[2] Western EcoSystems Technology, Inc., Cheyenne, Wyoming, U.S.A.

E-mail for correspondence: jennifer.brown@canterbury.ac.nz

Abstract: Ranked set sampling is a design that can increase sample efficiency by forcing sample coverage over the range of response values in the population. Balanced acceptance sampling (BAS) has a similar aim, with sample efficiency coming from ensuring coverage, where coverage is considered in terms of spatial spread over the sample space. Here we describe how ranked set sampling techniques can be used with BAS. We show how the two designs can be integrated and discuss the potential advantages of such an approach.

Keywords: BAS; environmental sampling; Halton sequence; ranked set sampling; spatial balance.

1 Ranked Set Sampling

Ranked set sampling (RSS) was first proposed by McIntyre (1952) for estimating the average yield of an arable crop in an agricultural field trial. Measuring yield for each field plot was time consuming because it involved removing, and then weighing, the crop. With some experience, he found it was relatively quick to estimate, by eye, the yield to the extent that a group of plots could be ranked in estimated yield order. McIntyre proposed the design, later named RSS (Halls and Dell 1966), where a subset of population units are first ranked by an auxiliary variable. The auxiliary variable can be a quick estimate of the response variable, as in the example described here, or some other variable known to be correlated to the response variable. Then, on the basis on these rankings, a sample is drawn from the ranked units.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), Guimarães, Portugal, 7–12 July 2019.

The RSS process for a sample of $n = km$ units is as follows:

1. Draw a simple random sample of k^2 units from the population, and randomly allocate the sampled units to k sets of k units. This gives k independent samples each of size k.
2. For set 1, rank order the units using an auxiliary variable and let $X[1]$ denote the unit with the lowest ranking. Include $X[1]$ in the ranked set sample and measure its response value. For the remaining $k-1$ units in the set, response values are not measured.
3. Repeat the step above for set 2, but now the unit with the second lowest ranking, $X[2]$, has its response value measured.
4. Repeat for set 3, and so on until the kth set has been ranked and the unit with the highest rank, $X[k]$, has been measured.

The steps above will result in k measured units, $X[1], \ldots, X[k]$, and represents one cycle. To obtain a RSS of size $n = km$, the entire process is repeated for m independent cycles. The process described here is called a balanced RSS and estimators for this design can be found in Wolfe (2010).

2 Balanced Acceptance Sampling

Balanced acceptance sampling (BAS) (Robertson et al. 2013) is a spatially balanced sampling design that spreads sample locations evenly over the sample space. Stevens and Olsen (2004) introduced spatially balanced sampling and defined spatial balance using the Voronoi tessellation of a sample. The closer the total inclusion probability in each Voronoi polygon is to one, the better the spatial balance. Spatially balanced designs are known to be efficient when surveying natural resources because nearby locations tend to be similar (Stevens and Olsen 2004; Robertson et al. 2013).

BAS uses a quasi-random number sequence, called a random-start Halton sequence (Halton 1960), to draw its sample. For ease of discussion, this article considers drawing sampling locations (points) from an areal resource $\Omega \subset [0,1]^2$ with $\lambda(\Omega) > 0$, where λ is the Lebesgue measure. The random-start Halton sequence $\{\mathbf{x}_j\}_{j=1}^{\infty}$ in $[0,1)^2$ is defined as follows. The ith coordinate of the jth point in this sequence is (Robertson et al. 2017)

$$x_j^{(i)} = \sum_{p=0}^{\infty} \left\{ \left\lfloor \frac{u_i + j}{b_i^p} \right\rfloor \bmod b_i \right\} \frac{1}{b_i^{p+1}},$$

where u_i is a random non-negative integer, $b_1 = 2$, $b_2 = 3$ and $\lfloor \cdot \rfloor$ is the floor function. The random-start Halton sequence is

$$\{\mathbf{x}_j\}_{j=1}^{\infty} = \left\{ (x_j^{(1)}, x_j^{(2)}) \right\}_{j=1}^{\infty}, \tag{1}$$

and setting $u_1 = u_2 = 0$ gives the classical Halton sequence (Halton 1960).

An equal probability BAS sample of size n is simply the first n points from (1) that fall within Ω. Unequal inclusion density functions can also be utilized with an acceptance rejection sampling technique (Robertson et al. 2013). BAS is conceptually simple, computationally efficient and is particularly useful for over-sampling because any sub-sequence of consecutive points from (1) is also spatially balanced (Robertson et al. 2018).

3 Balanced Acceptance Sampling with Ranked Set Sampling

The appealing feature of RSS is that there is sample coverage over the response variable, while with BAS sample coverage is over the sample space. Combining these two ideas means that BAS can potentially be further extended by integrating auxiliary information into the design through RSS.

Consider drawing a balanced RSS of $n = km$ points from Ω. First, draw an equal probability BAS over-sample of $n_0 = k^2 m$ points from Ω using (1) such that $\mathbf{x}_1 \subset \Omega$. For notational convenience, let $H = \{\mathbf{x}_j\}_{j=1}^{n_0}$ be the first n_0 (order preserved) points in (1) that are also in Ω. Take the first k^2 points and put them into k ordered sets of k points:

$$
\begin{array}{lllll}
\text{Set 1} & \{\mathbf{x}_1, & \mathbf{x}_2, & \ldots, & \mathbf{x}_k\} \\
\text{Set 2} & \{\mathbf{x}_{k+1}, & \mathbf{x}_{k+2}, & \ldots, & \mathbf{x}_{2k}\} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Set k} & \{\mathbf{x}_{(k-1)k+1}, & \mathbf{x}_{(k-1)k+2}, & \ldots, & \mathbf{x}_{k^2}\}.
\end{array}
$$

To remove ranking dependency between sets, let σ be a vector whose elements are a random permutation of $\{1, 2, \ldots, k\}$. For set 1, rank order the units using an auxiliary variable and let $X[\sigma(1)]$ denote the unit with the $\sigma(1)$ lowest ranking. Include $X[\sigma(1)]$ in the ranked set sample and measure its response value. Repeat the previous step for set 2, but now the unit with the $\sigma(2)$ lowest ranking, $X[\sigma(2)]$, has its response value measured. Repeat for set 3, and so on until the kth set has been ranked and $X[\sigma(k)]$ has been measured. This gives k measured units, $X[1], \ldots, X[k]$, and represents one cycle. To obtain the RSS of size $n = km$, the entire process is repeated for the next batch of points, $\{\mathbf{x}_{k^2+1}, \ldots, \mathbf{x}_{2k^2}\}$, and so on until the mth cycle.

The sampling strategy proposed here is highly structured. Each set of k points, each cycle of k^2 points, and the over-sample of $k^2 m$ points are all bona fide BAS samples (Robertson et al. 2018). Hence, each stage of the design draws spatially balanced samples, but the measured points may not be as spatially balanced as BAS alone. The idea here is to force spatial spread over the sample space at each stage of the design to promote sample coverage over the auxiliary variable. Then, based on this auxiliary information, ranked set sampling forces sample converge over the response to potentially increase the precision of BAS.

4 Preliminary Results and Discussion

Preliminary results using the test populations in Robertson et al. (2018) with perfect rankings show that BAS's simulated precision can be increased using RSS. Clearly perfect rankings may not be possible and there are sample costs to consider, but if there were a cost differential between a low-cost auxiliary variable and the higher cost response variable then these results are promising. This design is also appealing because it does not require auxiliary information a priori for every population unit. Other approaches, for example, unequal probability designs, may require accurate auxiliary information to define each unit's inclusion probability before the sample is drawn, which can be infeasible in practice.

References

Halls, L.K. and Dell, T.R. (1966). Trial of ranked-set sampling for forage yields. *Forest Science*, **12**, 22 – 26.

Halton, J.H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, **2**, 84 – 90.

McIntyre, G.A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, **3**, 385 – 390.

Robertson, B.L., Brown, J.A., McDonald, T., and Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, **3**, 776 – 784.

Robertson, B.L., McDonald, T., Price, C.J., and Brown, J.A. (2017). A modification of balanced acceptance sampling. *Statistics and Probability Letters*, **129**, 107 – 112.

Robertson, B.L., McDonald, T., Price, C.J., and Brown, J.A. (2018). Halton iterative partitioning: spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, **3**, 305 – 323.

Stevens, D.L., Jr. and Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, 262 – 278.

Wolfe, D.A (2010). Ranked set sampling. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 460 – 466.